Cross-Representation Transferability of Adversarial Attacks: From Spectrograms to Audio Waveforms

Karl Michel Koerich*, Mohammad Esmailpour†, Sajjad Abdoli†, Alceu de S. Britto Jr.‡
and Alessandro L. Koerich†

*McGill University, Montreal, QC, Canada
Email: karl.koerich@mail.mcgill.ca
†École de Technologie Superiéure, Université du Québec, Montréal, QC, Canada
Email: mohammad.esmailpour.1@ens.etsmtl.ca, sajjad.abdoli.1@ens.etsmtl.ca, alessandro.koerich@etsmtl.ca
‡Pontifical Catholic University of Paraná, Curitiba, PR, Brazil
Email: alceu@ppgia.pucpr.br

***Abstract*—This paper shows the susceptibility of spectrogram-based audio classifiers to adversarial attacks and the transferability of such attacks to audio waveforms. Some commonly used adversarial attacks to images have been applied to Mel-frequency and short-time Fourier transform spectrograms, and such perturbed spectrograms are able to fool a 2D convolutional neural network (CNN). Such attacks produce perturbed spectrograms that are visually imperceptible by humans. Furthermore, the audio waveforms reconstructed from the perturbed spectrograms are also able to fool a 1D CNN trained on the original audio. Experimental results on a dataset of western music have shown that the 2D CNN achieves up to 81.87% of mean accuracy on legitimate examples and such performance drops to 12.09% on adversarial examples. Likewise, the 1D CNN achieves up to 78.29% of mean accuracy on original audio samples and such performance drops to 27.91% on adversarial audio waveforms reconstructed from the perturbed spectrograms.**

***Index Terms*—Adversarial audio attacks, transferability, audio reconstruction.**

I. Introduction

Music genre classification is a challenging task for humans [1]–[6] due to the subjectivity and unclear boundaries between genres, and the uniqueness of musicians and artists. Yet, well classifying music is of great interest to many researchers and companies in the entertainment and arts industry. In the last years, convolutional neural networks (CNNs) became increasingly popular due to their high accuracy and performance on image datasets. Therefore, the focus in academia has been on 2D CNNs. When it comes to audio and music processing, CNNs have had a significant impact on several tasks such as automatic music tagging [7], video clip classification based on audio information [8], speaker identification [9], environmental sound classification [10], [11] and music genre classification [12]–[14].

Even if audio is a 1D signal, it is a common practice to use 2D representations, like spectrograms, when training machine learning models. Due to their ability to model the human peripheral auditory system [15], Mel-frequency cepstrum coeficient (MFCC) features are currently used for several audio processing tasks such as music genre classification [16]. Mel-frequency (MF) spectrograms, which use Mel-frequency filter banks to represent the short-term power spectrum of audio on the Mel-scale of frequency, are one of the most preferred input types for music information retrieval [17]. One of the main advantages of using 2D representations is that they summarize high dimensional waveforms into compact time-frequency representations while audio signals alone are noisier [18]. Regardless of the type of spectrogram; as they are 2D representations of audio signals, they can be treated as images. Therefore, this opens up the opportunity to benefit from the recent advances of deep neural networks in computer vision.

Recent works have exploited the capability of CNNs to learn representations directly from spectrograms. Boddapati *et al.* [19] use short-time Fourier transform (STFT), MFCC and Cross Recurrence Plot (CRP) spectrograms with two different 2D CNN architectures (AlexNet and GoogLeNet) to classify 2D representations of environmental sounds. Lee *et al.* [20] use 2D CNNs with MF spectrograms as input for music auto-tagging. Pons *et al.* [21] use 40 bands MF spectrograms to experiment with musically motivated CNNs and try understanding what CNNs learn from particular datasets. Pons *et al.* [12] use MF spectrograms as input for a randomly weighted CNN for music audio classification. Oramas *et al.* [22] use constant-Q transform (CQT) spectrograms in their audio-based approach for multi-label music genre classification.

Despite all advantages, it has been shown that approaches based on 2D representations are susceptible to adversarial attacks, which can easily fool these models and raise safety concerns. Esmailpour *et al.* [23] have shown that the majority of state-of-the-art approaches for audio classification relying on 2D CNNs can be easily deceived, with fooling rates higher than 90% and high confidence. 1D CNNs can also be easily fooled by adversarial attacks. Abdoli *et al.* [24] demonstrated the existence of universal adversarial perturbations that can fool several audio processing architectures with attack success rates between 91.1% and 74.7%, and signal-to-noise ratio (SNR) between 15.70dB and 19.62dB. Du *et al.* [25] proposed a method based on Particle Swarm Optimization (PSO) for generating adversarial audio for end-to-end audio systems. They evaluated their attacks on a range of applications

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

like speech command recognition, speaker recognition, sound event detection and music genre classification. The proposed attack achieved a success rate of 89.30% on a 1D CNN and 91.20% on a convolutional recurrent neural network with SNR of 15.39dB and 17.24dB respectively. However, the low SNRs indicate that the adversarial perturbations are audible and can be easily perceived by a listener.

Even if few works have already studied adversarial attacks on both 1D and 2D CNNs [14], [23]–[25], none of them have evaluated the transferability of such adversarial attacks across representations, in particular from 2D to 1D. Generating 1D adversarial attacks is much more time-consuming than generating 2D attacks due to the high dimensionality of audio signals [23], as it requires computing a similarity measure such as the ℓ_2-norm between legitimate and crafted examples as a part of an adversarial optimization criterion. Therefore, it might be advantageous to generate 2D perturbations and transfer it to audio waveforms.

The main contributions of this paper are: (i) we show that the most effective adversarial attacks for images can also attack spectrograms generated from music, which reveals the vulnerability of 2D CNNs to non-specific attacks; (ii) we show that perturbed spectrograms can be used to reconstruct audio signals that are perceptually similar to the original audio, with an SNR value of about 20dB; (iii) we show that the audio waveforms reconstructed from the perturbed spectrograms can also fool a 1D CNN trained on the original audio with high confidence.

This paper is organized as follows. Section II presents the baseline 1D and 2D CNN architectures for music genre classification. Section III presents a description of the adversarial attacks and the reconstruction process of the audio waveforms from spectrograms. Section IV presents the dataset used, the experimental protocol, and the experimental results. We compare the performance of the 2D CNN model when using MF and STFT spectrograms, and the vulnerability of such a model to adversarial examples. We also evaluate the transferability of adversarial perturbations from 2D representations to the audio waveform and the susceptibility of a 1D architecture to such a transferred attack. Finally, the conclusions and perspectives of future work are presented in the last section.

II. 1D and 2D CNN Architectures

The aim of the proposed architecture is to deal with 2D representations of audio signals of various lengths, learning meaningful representations directly from spectrograms. First, we split each audio waveform into fixed-length segments using a sliding window of appropriate width. The window width depends mainly on the signal sampling rate, which in the case of the music dataset evaluated in this paper is 22,050 Hz. In our approach, we use a window of five seconds (110,250 samples) because such a length provides the best trade-off between the number of segments and accuracy. Furthermore, there is also a certain percentage of overlapping between successive audio segments, which aim is to maximize the use of information. In our approach we use 75% overlapping because such percentage of overlap provides the best trade-off between the number of segments and accuracy. Furthermore, the overlapping can be viewed as some sort of data augmentation since it naturally increases the number of samples due to the fact that some parts of the audio signal are reused. Fig. 1 summarizes the process of splitting the audio file into appropriate segments by the sliding window, then having the signals converted to spectrograms, which are then used as input to the 2D CNN. A similar process is used with the 1D CNN except that the input of the CNN is audio segments, as illustrated in Fig. 2.

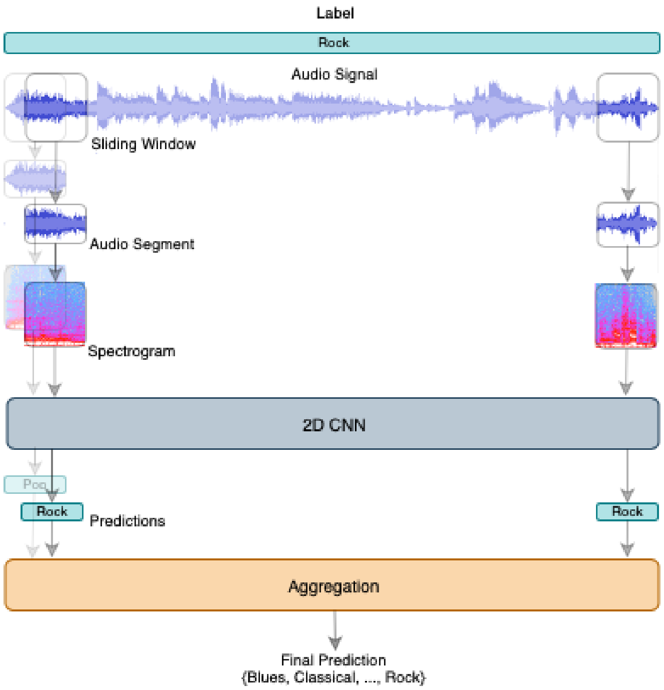

Fig. 1. Overview of the proposed approach for audio signal segmentation by a sliding window, transformation of audio segments into spectrograms, prediction by a 2D CNN and aggregation of prediction on the segments to come up to a final prediction.

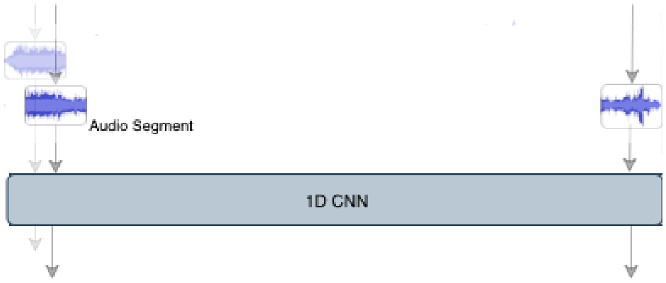

Fig. 2. Overview of the changes in Fig. 1 for the 1D approach. The 1D CNN receives the waveform segments instead of spectrograms and provides the predictions on segments for aggregation.

Moreover, the length of the original audio (before being split) has a direct impact on the number of samples being tested and trained, which may impact the computational cost of the model. The GTZAN dataset has a sampling rate of 22,050 Hz and all original audio files are 30 seconds long.

STFT and MF are the main approaches for producing spectrograms for music signals. To generate the STFT and MF spectrograms, we use a fast Fourier transform (FFT) window of length 512 and 256 samples between successive frames. For the MF spectrograms we use 64 Mel filters. Therefore, the spectrogram dimension becomes 431×257 for STFT, and 431×64 for MF. The architecture of the 2D CNN is the same for both input formats and it has two convolutional layers (CL), followed by a maxpooling layer to reduce the complexity of the network, followed by another two CLs and another maxpooling layer, which is connected to a fully connected (FC) layer followed by an output layer. ReLU is used as activation function in all layers but the last which uses softmax. The weights of all of layers are initialized randomly. Batch normalization is included between CLs and a dropout of 0.4 is used after the FC layer. The breakdown of the 2D CNN is presented in Table I.

Several 1D CNN architectures have been proposed to deal directly on the audio waveforms [24]. The 1D CNN model receives 5-second audio segments as input and it has five one-dimensional CLs, where the first CL employs a Gammatone filter-bank. This layer is kept frozen during the training process. Gammatone filters are used to decompose the input audio signal to appropriate frequency bands. The output of the last CL is used as input to one FC layer followed by an output layer. Leaky ReLU is used as activation function in all layers except in the last which uses softmax. The weights of all of the layers are initialized randomly. This model was proposed by Abdoli *et al.* [11] for environmental sound classification. Batch normalization is included between CLs and a dropout of 0.4 is used after the FC layer. The breakdown of the 1D CNN is presented in Table II.

TABLE I
ARCHITECTURE OF THE PROPOSED 2D CNN

Layer	# of	Filter	Stride	Output Shape	
Type	Filters	Size		STFT	MF
Input	-	-	-	257, 431, 1	64, 431, 1
Conv2D	32	3, 3	1, 1	255, 429, 32	62, 429, 32
Conv2D	32	3, 3	1, 1	253, 427, 32	60, 427, 32
MaxPool	-	-	2, 2	126, 213, 32	30, 213, 32
Conv2D	64	3, 3	1, 1	124, 211, 64	28, 211, 64
Conv2D	64	3, 3	1, 1	122, 209, 64	26, 209, 64
MaxPool	-	-	2, 2	60, 103, 64	12, 103, 64
Dense	-	-	-	1 024	1 024
Output	-	-	-	10	10

During the classification step, since the input audio waveform is split into several segments, we need to aggregate the 2D CNN predictions to come up to a final decision on the input audio, as illustrated in Fig. 1. For such an aim, we used majority vote and the sum rule [11]. When there are K classes, we generate K values and we choose the class which has the maximum value as our final prediction. A similar process is used to aggregate the predictions of the 1D CNN, as illustrated in Fig. 2.

TABLE II
ARCHITECTURE OF THE PROPOSED 1D CNN

Layer Type	# of Filters	Filter Size	Stride	Output Shape
Input	-	-	-	110 250
Conv1D Gamma	32	512	1	109 739
AvgPool	-	-	8	13 717
Conv1D	16	256	2	6 731
AvgPool	-	-	8	841
Conv1D	32	64	2	389
Conv1D	64	32	2	179
Conv1D	128	16	2	82
MaxPool	-	-	2	41
Dense	-	-	-	256
Output	-	-	-	10

III. ADVERSARIAL ATTACKS AND AUDIO RECONSTRUCTION

Adversarial attacks can be considered as small crafted perturbations that, when intentionally added to a legitimate example, lead machine learning models to misbehave [26]. Considering x as a legitimate example, then an adversarial example x' can be crafted in such a way that:

$$x \approx x', \qquad f^*(x) \neq f^*(x') \tag{1}$$

where f^* is the post-activation function. Assuming that x is a spectrogram, the crafted x' should be unrecognizable by human visual system.

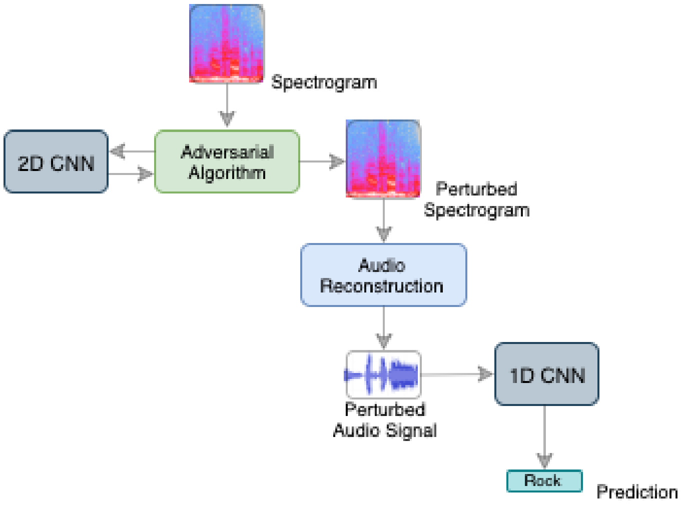

Fig. 3. Overview of the 2D adversarial attack which relies on the 2D CNN model and training dataset (spectrograms) to produce adversarial spectrograms. The audio waveform is reconstructed from the perturbed spectrogram and used to fool the 1D CNN model.

Among the several algorithms for generating x', the Fast Gradient Sign Method (FGSM) [27] was one of the first attacks, which still remains one of the most effective adversarial attacks. Goodfellow *et al.* [27] introduced the FGSM for generating simple adversarial samples. The method consists of adding to the legitimate example $\mathbf{x}$ an imperceptibly small perturbation that is equal to the product of a small constant ϵ and the sign of the gradient of the cost function J for the model parameter $\mathbf{w}$ with respect to the input $\mathbf{x}$ and the true label y, as shown in (2).

$$\mathbf{x}' = \mathbf{x} + \epsilon \cdot \mathrm{sign}(\nabla_{\mathbf{x}} J(\mathbf{w}, \mathbf{x}, y)) \tag{2}$$

The resulting adversarial example $\mathbf{x}'$ carries a small perturbation that cannot be seen by the human eye and effortlessly deceives 2D CNNs and other non-deep architectures [23].

Kurakin *et al.* [28] introduced a straightforward way of extending the FGSM method by applying it multiple times with a small step size. Known as the Basic Iterative Method (BIM), this adversarial attack is also able to fool complex 2D CNNs. As illustrated in the upper part of Fig. 3, both FGSM and BIM are white-box adversarial attacks which means that both the trained 2D CNN model and the training dataset should be accessible to the adversarial algorithms to fetch its gradient information and generate the adversarial input $\mathbf{x}'$ with unrecognizable differences to the legitimate input $\mathbf{x}$. The perturbed spectrogram can perhaps make a 2D CNN predict a wrong label with high confidence. The lack of robustness of 2D CNNs to these two attacks was also observed for the task of environmental sound classification [23], [29].

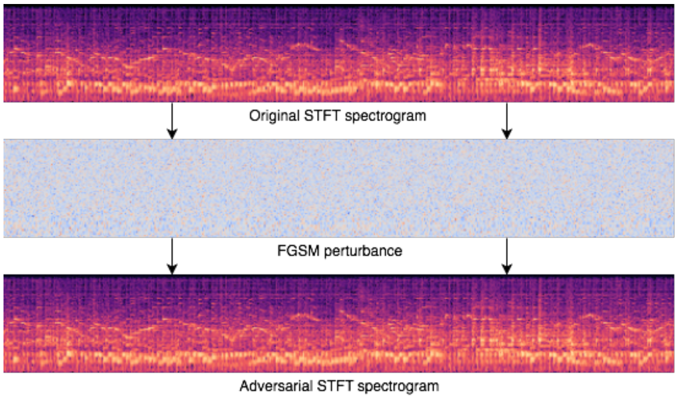

Fig. 4. The original STFT spectrogram and the perturbation produced by the FGSM attack to generate an adversarial STFT spectrogram.

Fig. 4 shows an example of a legitimate STFT spectrogram attacked with a perturbation produced by the FGSM attack. Such a figure summarizes what happens with attacked STFT and MF spectrograms: the difference between the adversarial and legitimate spectrograms is imperceptible to the human visual system. This aspect is very important, otherwise the perturbation could not be considered as an adversarial one. However, what happens with the perturbed signal in its original representation space, that is, audio waveform? Will such a perturbation remain unperceived by the human auditory system?

A. Audio Reconstruction

The main contribution of this paper is to evaluate if after performing adversarial attacks to the 2D representations and successfully fooling the 2D CNN, such attacks could be transferred to the audio waveform. If so, the perturbations should not be perceptible by human auditory system on the audio after reconstructing it from the perturbed spectrogram. Besides evaluating the transferability of the adversarial attacks across representations (from 2D to 1D), we also want to evaluate if such an reconstructed adversarial audio is able to fool an 1D CNN model which accomplishes the same task of the 2D CNN. For such an aim, we need to reconstruct the audio signals from the spectrograms.

The phase of the audio signal is fundamental to an accurate reconstruction of the signal. However, both the STFT and MF spectrograms do not contain information about the exact, or even approximate, phase of the signal that it represents. Therefore, it is not possible to reverse the process without distorting the reconstructed signal.

To circumvent this problem, we can retain the phase information separately and, depending on the type of spectrogram, we can use such information to reconstruct the original audio signal, which is the case of the STFT spectrogram. This avoids the introduction of distortions to the reconstructed audio signal, which may tamper or even mask the adversarial perturbations embedded into the reconstructed adversarial audio examples. On the other hand, to reconstruct the audio signals from the MF spectrograms, it is necessary to estimate the unknown phases iteratively using the Griffin-Lim (GL) algorithm [30].

IV. Experimental Results

The proposed 2D CNN for music genre classification was evaluated on the GTZAN dataset. This dataset consists of 1,000 30-second audio clips evenly distributed into 10 classes: Blues, Classical, Country, Disco, Hip-Hop, Jazz, Metal, Pop, Reggae and Rock. The audio samples were collected from a variety of sources in order to represent a variety of recording conditions [31]. Even if the GTZAN dataset has several known problems with its integrity, such as replications, mislabeling, and distortions [14], [32], this does not affect our experiments since our aim is not assessing the accuracy of CNN models, but their vulnerabilities against adversarial attacks.

The 1,000 audio files were shuffled and divided into three folds with 340, 330 and 330 samples, respectively. Fold 1 contains 34 tracks of each of the 10 genres; Folds 2 and 3 contain 33 tracks of each genre. Every track of every fold was split into 21 short segments according to the sliding window described in Fig. 1. The next step was to generate spectrograms of all 21,000 audio segments. The model was first accessed with MF spectrograms and then STFT spectrograms. Two folds were used for training and 20% of the training set was used for validation. The third fold was used for testing. After predicting the music genre for each segment on the testing set, the predictions for all 21 windows belonging to the same song are aggregated with majority vote and sum rule to determine the final genre prediction for the whole track. Each network was trained up to 100 epochs with batch sizes of 50 samples and early stopping. The model's performance for each input type is presented in Table III. The performance reported in Table III is far from the best performance already achieved for such a dataset which is about 90% [33]. On the other hand, the results are similar to Kereliuk *et al.* [14] that used two different 2D CNN architectures and achieved 81.20% of mean accuracy for the same dataset. Furthermore, we did not attempt to fully optimize the performance of the 2D CNN to reduce the risk of overfitting, as it can potentially increase the susceptibility of CNNs to adversarial attacks.

TABLE III
MEAN ACCURACY AND STANDARD DEVIATION OF THE 2D CNN FOR STFT AND MF SPECTROGRAMS.

	Mean Accuracy ± SD (%)		
Input	**Segments**	**MV Aggregation**	**SR Aggregation**
STFT	67.09±1.49	75.84±1.92	75.64±2.09
MF	75.29±2.50	81.87±2.49	81.37±2.20

MV: Majority Voting; SR: Sum Rule; SD: Standard Deviation

Table IV shows the performance achieved by the 1D CNN on original audio waveforms as well as on the waveforms reconstructed from the STFT spectrogram with the original phase information. The performance achieved on the reconstructed audio is slight better than that achieved on the original audio. This is a clear indication that the reconstruction process is accurate and it does not insert spurious noise to the signal.

TABLE IV
MEAN ACCURACY AND STANDARD DEVIATION OF THE 1D CNN FOR ORIGINAL AUDIO AND AUDIO SIGNAL RECONSTRUCTED FROM STFT SPECTROGRAM AND PHASE INFORMATION.

	Mean Accuracy ± SD (%)		
Input	**Segments**	**MV Aggregation**	**SR Aggregation**
Original	70.82±0.20	77.69±0.81	77.99±1.04
STFT Recons	71.87±0.67	78.49±0.26	78.29±0.72

MV: Majority Voting; SR: Sum Rule; SD: Standard Deviation

Table V shows the results of the FGSM and BIM attacks against the 2D CNN. We evaluate the mean accuracy achieved by the model on the perturbed spectrograms. For both STFT and MF spectrograms the BIM attack is more successful due to its iterative nature. For instance, considering the best result of Table III, the mean accuracy for STFT spectrogram drops from 75.84% to 11.58% and from 81.87% to 12.09% for the MF spectrogram.

Table VI shows the performance of the 1D CNN on the adversarial audio samples generated by the reconstruction of the STFT spectrogram perturbed with FGSM and BIM. The mean accuracy drops from 77.99% on the original audio to 27.91% on the examples attacked by FGSM. We did not evaluate audio examples reconstructed from MF spectrograms because even the reconstruction of legitimate examples is very noisy due to the approximate phase estimation by the GL algorithm.

TABLE V
MEAN ACCURACY AND STANDARD DEVIATION FOR THE 2D CNN AFTER ATTACKING STFT AND MF SPECTROGRAMS WITH FGSM AND BIM ADVERSARIAL ATTACKS.

		Mean Accuracy ± SD (%)		
Input	**Attack**	**Segments**	**MV Aggreg.**	**SR Aggreg.**
STFT	FGSM	17.40±1.20	17.35±0.97	16.76±1.19
	BIM	13.26±1.20	11.58±0.97	9.89±1.19
MF	FGSM	21.52±1.80	19.32±1.46	19.41±0.88
	BIM	15.38±1.99	12.09±2.01	12.49±1.88

MV: Majority Voting; SR: Sum Rule; SD: Standard Deviation

TABLE VI
MEAN ACCURACY AND STANDARD DEVIATION FOR THE 1D CNN AFTER RECONSTRUCTING AUDIO FROM STFT SPECTROGRAMS ATTACKED WITH FGSM AND BIM ADVERSARIAL ATTACKS.

		Mean Accuracy ± SD (%)		
Input	**Attack**	**Segments**	**MV Aggreg.**	**SR Aggreg.**
STFT Recons	FGSM	26.45±1.09	28.00±0.98	27.91±0.85
	BIM	30.76±2.03	33.85±3.45	33.35±3.67

MV: Majority Voting; SR: Sum Rule; SD: Standard Deviation

Finally, we need to evaluate if the adversarial perturbation added to the spectrogram will remain unrecognizable by human auditory system when the audio waveform is reconstructed from the adversarial spectrogram. For such an aim we conducted two experiments: (i) a quantitative experiment using signal-to-noise ratio (SNR) as a metric to measure the level of the noise with respect to the original signal; (ii) a qualitative listening experiment with expert and non-expert listeners.

SNR has been used by previous works to evaluate the quality of the generated adversarial audio by measuring the level of the perturbation on the signal after adding the perturbations [14], [24], [25] and it is defined as:

$$\mathrm{SNR}_{\mathrm{dB}}(\mathbf{x}^r, \mathbf{v}) = 20.\log_{10}\frac{P(\mathbf{x}^r)}{P(\mathbf{v})}, \tag{3}$$

where $\mathbf{x}^r$ denotes the audio reconstructed from the legitimate spectrogram and $\mathbf{v}$ denotes the adversarial noise. $P(.)$ is the power of the signal or noise, which is defined as:

$$P(\mathbf{x}) = \sqrt{\frac{1}{N}\sum_{n=1}^{N} x_n^2}, \tag{4}$$

where x_n denotes the n-th component of the signal $\mathbf{x}$. A high SNR indicates that a low level of noise is added to the audio signal by the adversarial perturbation.

According to Du *et al.* [25], the noise is imperceptible when $\mathrm{SNR}_{\mathrm{dB}}$ is equal or greater than 20dB. This is also supported by the experiments carried out by Abdoli *et al.* [11] for environmental sound classification[1]. Table VII shows the

[1] See: https://sajabdoli.netlify.com/publication/uap/ for some audio samples

mean SNR achieved on the reconstructed audio from the spectrograms attacked by FGSM and BIM, and None refers to the SNR between the original and reconstructed audio. In the case of STFT spectrograms, which uses the phase information in the reconstruction process, the reconstructed audio is more accurate and it becomes equivalent to the original audio (SNR $>$ 90dB). On the other hand, MF spectrograms rely on the phase estimation using Griffin-Lim's method, thus the reconstructed audio is quite noisy (SNR $<$ 20dB). Therefore, the SNR for the audio reconstructed from MF spectrograms is quite misleading because the noisy reconstruction also masks the adversarial perturbation.

TABLE VII
MEAN SNR AND STANDARD DEVIATION (SD) COMPUTED BY (4) FOR THE AUDIO RECONSTRUCTED FROM THE PERTURBED STFT AND MF SPECTROGRAMS BY THE FGSM AND BIM ATTACKS.

Input	**Attack**	$\mathbf{SNR_{dB}(x, v)} \pm$ **SD**
Audio Reconstructed (STFT)	None	>90
	FGSM	14.62±5.93
	BIM	20.19±5.95
Audio Reconstructed (MF)	None	8.71±11.77
	FGSM	42.69±9.45
	BIM	44.21±9.90

The qualitative experiment was very limited and it was conducted with only four listeners using the speakers of a desktop computer. Pairs of legitimate and adversarial audio of the same song in a random order were presented to listeners and they pointed out whether they perceive or not differences in the audio samples. The results are disappointing because in average, listeners have noticed audible difference in all audio pairs. Nevertheless, most of the listeners referred to a "small background noise" which is not even close to change the perception of the listeners about the musical genre. Several examples of legitimate and adversarial audio samples are available to readers[2].

V. CONCLUSION

This paper presented a 2D CNN for music genre classification and evaluated it with two 2D representations: MF spectrograms and STFT spectrograms. The proposed approach learns from spectrograms of audio segments and performs relatively well compared to the state-of-the-art on the GTZAN dataset. The proposed 2D CNN achieved 81.87% and 75.84% of mean accuracy for MF and STFT spectrograms, respectively. Even if spectrograms seem to be advantageous to model in a compact and informative way the spectrum of frequencies of an audio signal as it varies with time, such 2D representations and 2D CNN models may not be the safest ones when it comes to robustness against adversarial attacks.

In this paper we have shown that adversarial attacks designed to regular images can also harm 2D representation of audio signals since the perturbations remain visually imperceptible on spectrogram images. FGSM and BIM attacks successfully fooled the 2D CNN in the task of music genre classification. These adversarial examples produced using 2D representation and a 2D CNN model were then transferred to audio waveforms. The audio waveforms produced were tested against a 1D CNN model and successfully fooled it. Therefore, we have shown the transferability of adversarial perturbations across representations. These results expose the vulnerability of both 1D and 2D CNN architectures to adversarial attacks.

The audio signals reconstructed from STFT spectrograms using the phase information have a very high SNR and the adversarial audio reconstructed from such spectrograms have SNRs between 14dB and 20dB. Nevertheless, the reconstructed adversarial audio examples are distinguishable from their legitimate counterparts to human perception according to the outcome of the qualitative evaluation. Even if the audio signals reconstructed from MF spectrograms do not have SNRs as high as those achieved using STFT spectrograms, the adversarial audio reconstructed from MF spectrograms have SNR greater than 20dB, which is mostly due to the noise introduced in the reconstruction process and not the adversarial perturbations.

Future work related to this paper includes adding acoustic perceptual considerations to eliminate spectral components in a way that no difference is perceived by the listeners. Also, since we showed the transferability of adversarial perturbations, future work includes the possibility of creating black-box adversarial attacks to 1D CNN models by using auxiliary 2D CNN models to produce adversarial images and then have them transferred to audio waveforms.

[2] https://github.com/karlmiko/ijcnn2020

REFERENCES

[1] C. H. L. Costa, J. D. Valle, and A L. Koerich, "Automatic classification of audio data," in *IEEE Intl Conf on Systems, Man and Cybernetics*, 2004, pp. 562–567.

[2] C. N. Silla, A. L. Koerich, and C. A. A. Kaestner, "A machine learning approach to automatic music genre classification," *Journal of the Brazilian Computer Society*, vol. 14, no. 3, pp. 7–18, 2008.

[3] T. Lidy, C. N. Silla Jr., O. Cornelis, F. Gouyon, A. Rauber, C. A. A. Kaestner, and A. L. Koerich, "On the suitability of state-of-the-art music information retrieval methods for analyzing, categorizing and accessing non-western and ethnic music collections," *Signal Processing*, vol. 90, no. 4, pp. 1032–1048, 2010.

[4] Y. M. G. Costa, L. E. S. Oliveira, A. L. Koerich, and F. Gouyon, "Music genre recognition using spectrograms," in *18th Intl Conf on Systems, Signals and Image Processing*, 2011, pp. 151–154.

[5] Y. M. G. Costa, L. E. S. Oliveira, A. L. Koerich, F. Gouyon, and J. G. Martins, "Music genre classification using LBP textural features," *Signal Processing*, vol. 92, no. 11, pp. 2723–2737, 2012.

[6] A. L. Koerich, "Improving the reliability of music genre classification using rejection and verification," in *14th Intl Society for Music Information Retrieval Conference (ISMIR)*, 2013, pp. 511–516.

[7] S. Dieleman and B. Schrauwen, "End-to-end learning for music audio," in *IEEE Intl Conf on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 6964–6968.

[8] S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. J. Weiss, and K. Wilson, "CNN architectures for large-scale audio classification," in *Intl Conf on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 131–135.

[9] M. Ravanelli and Y. Bengio, "Speaker recognition from raw waveform with sincnet," in *IEEE Spoken Language Technology Workshop (SLT)*, 2018, pp. 1021–1028.

[10] M. Esmaeilpour, P. Cardinal, and A. L. Koerich, "Unsupervised feature learning for environmental sound classification using weighted cycle-consistent generative adversarial network," *Applied Soft Computing*, vol. 86, pp. 105912, 2020.

[11] S. Abdoli, P. Cardinal, and A. L. Koerich, "End-to-end environmental sound classification using a 1D convolutional neural network," *Expert Syst. Appl.*, vol. 136, pp. 252–263, 2019.

[12] J. Pons and X. Serra, "Randomly weighted cnns for (Music) audio classification," in *IEEE Intl Conf on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 336–340.

[13] K. Choi, G. Fazekas, M. Sandler, and K. Cho, "Convolutional recurrent neural networks for music classification," in *IEEE Intl Conf on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 2392–2396.

[14] C. Kereliuk, B. L. Sturm, and J. Larsen, "Deep learning and music adversaries," *IEEE Trans Multimedia*, vol. 17, no. 11, pp. 2059–2071, 2015.

[15] V. Tiwari, "Mfcc and its applications in speaker recognition," *Intl Journal on Emerging Technologies*, vol. 1, no. 1, pp. 19–22, 2010.

[16] D. Iter, J. Huang, and M. Jermann, "Generating adversarial examples for speech recognition," *Stanford Technical Report*, 2017.

[17] S. Panwar, A. Das, M. Roopaei, and P. Rad, "A deep learning approach for mapping music genres," in *12th System of Systems Engineering Conf (SoSE)*, 2017, pp. 1–5.

[18] D. Stowell and M. D. Plumbley, "Automatic large-scale classification of bird sounds is strongly improved by unsupervised feature learning," *PeerJ*, vol. 2, pp. e488, 2014.

[19] V. Boddapati, A. Petef, J. Rasmusson, and L. Lundberg, "Classifying environmental sounds using image recognition networks," *Procedia computer science*, vol. 112, pp. 2048–2056, 2017.

[20] J. Lee and J. Nam, "Multi-level and multi-scale feature aggregation using pretrained convolutional neural networks for music auto-tagging," *IEEE Signal Processing Letters*, vol. 24, pp. 1208–1212, 2017.

[21] J. Pons and T. Lidy, "Experimenting with Musically Motivated Convolutional Neural Networks," in *14th Intl Workshop on Content-Based Multimedia Indexing*. 2016, IEEE.

[22] S. Oramas, O. Nieto, F. Barbieri, and X. Serra, "Multi-label music genre classification from audio, text, and images using deep features," in *Intl Conf on Music Information Retrieval (ISMIR)*, 2017, pp. 23–30.

[23] M. Esmaeilpour, P. Cardinal, and A. L. Koerich, "A robust approach for securing audio classification against adversarial attacks," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 2147–2159, 2020.

[24] S. Abdoli, L. G. Hafemann, J. Rony, I. Ben Ayed, P. Cardinal, and A. L. Koerich, "Universal adversarial audio perturbations," *CoRR*, vol. abs/1908.03173, 2019.

[25] T. Du, S. Ji, J. Li, Q. Gu, T. Wang, and R. Beyah, "Sirenattack: Generating adversarial audio for end-to-end acoustic systems," *arXiv preprint arXiv:1901.07846*, 2019.

[26] T.-W. Weng, H. Zhang, P.-Y. Chen, J. Yi, D. Su, Y. Gao, C.-J. Hsieh, and L. Daniel, "Soundnet: Learning sound representations from unlabeled video," in *6th Intl Conf Learning Representation*, 2018.

[27] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *3rd Intl Conf on Learning Representations*, 2015.

[28] A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," *arXiv preprint arXiv:1607.02533*, 2016.

[29] M. Esmailpour, P. Cardinal, and A. L. Koerich, "Detection of adversarial attacks and characterization of adversarial subspace," in *IEEE Intl Conf on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 3097–3101.

[30] D. Griffin and J. Lim, "Signal estimation from modified short-time Fourier transform," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 2, pp. 236–243, 1984.

[31] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 5, pp. 293–302, 2002.

[32] Bob L. Sturm, "An analysis of the GTZAN music genre dataset," in *2nd Intl ACM workshop on Music information retrieval with user-centered and multimodal strategies (MIRUM)*, 2012, p. 7.

[33] Y. Panagakis, C. Kotropoulos, and G. R. Arce, "Music genre classification using locality preserving non-negative tensor factorization and sparse representations," in *10th Intl Society for Music Information Retrieval Conference (ISMIR)*, 2009, pp. 249–254.